Best arm identification via Bayesian gap-based exploration

Matthew W. Hoffman[*] HOFFMANM@CS.UBC.CA
Bobak Shahriari[*] BSHAHR@CS.UBC.CA
Nando de Freitas NANDO@CS.UBC.CA
University of British Columbia; Vancouver, BC, Canada

Abstract

Bayesian approaches to optimization under bandit feedback have recently become quite popular in the machine learning community. Methods of this type have been found to have not only very good empirical performance, but also optimal theoretical regret bounds when analyzed from a frequentist perspective. In this work we study theoretical, methodological, and empirical aspects of the problem of best arm identification in stochastic multi-armed bandits from a Bayesian perspective. In particular, we introduce a Bayesian version of the gap-based method of (Gabillon et al., 2012). In the domain of sensor networks, with real traffic data, this approach shows significant gains in performance over both Bayesian cumulative regret techniques and frequentist simple regret methods.

1. Introduction

The problem of best arm identification in stochastic multi-armed bandit problems has recently received a great deal of theoretical attention (see Bubeck et al., 2009; Audibert et al., 2010). As in more standard multi-armed bandit settings, this problem revolves around a decision maker who must repeatedly take actions, i.e. by selecting an arm and observing a reward for pulling that arm; see for example (Berry & Fristedt, 1985; Cesa-Bianchi & Lugosi, 2006; Gittins et al., 2011) for extensive discussions. However, unlike more standard settings, the goal is not to maximize the cumulative sum of these observed rewards. Instead, the decision maker is allowed to interact with the bandit process during an *exploration* phase after which they are required to recommend a single arm; the decision maker is then judged only on the value of the single arm that is recommended.

An almost canonical example of the best arm identification problem, also known as *pure exploration*, is that of product testing. Take for example, a company considering different marketing strategies for their products. The company might consider presenting these different strategies to a subset of their potential customers. In this setting, the company is not necessarily interested in persuading those particular customers, but is instead concerned with the problem of finding the best strategy for selling products to their customer base at large. This initial, limited exploration phase serves as a proxy for the much larger set of customers. We can then ask the question of how best to query customers during this exploratory phase in order to have the highest probability of success in the *testing* phase when the marketing strategy is ultimately rolled out. In this same vein, the popular “A/B testing” framework, used for tailoring many of the design choices of modern web and mobile applications, can be seen as a problem of best arm identification (see Scott, 2010; Kohavi et al., 2009).

Another area of bandit research that has received a great deal of attention recently involves incorporating Bayesian methods. Although some of the earliest work on what would now be called bandit problems came from a Bayesian perspective (Thompson, 1933), the field has since become dominated by frequentist approaches based on regret minimization (Robbins, 1952; Lai & Robbins, 1985). See also the work of (Auer et al., 2002). In the past few years, however, Bayesian methods have noticed something of a resurgence, partially due to their great empirical performance (Chapelle & Li, 2012; Scott, 2010). More recent theoretical work has also shown that even while these methods take a Bayesian approach to modeling each arm, they still possess optimal, cumulative regret guarantees. See the work of (Kaufmann et al., 2012a) for bounds on a Bayesian version of the classical upper confidence bound (UCB) approach, as well as (Kaufmann et al.,

[*] Authors contributed equally.

2012b) for bounds on an approach based on the original work of Thompson. More recently, work of (Russo & Van Roy, 2013) has explicitly considered similar sample-based approaches with correlation between the arms.

We would also be remiss if we did not mention Bayesian optimization. Although a full review of this area is beyond the scope of this work, it does share a great deal of overlap with bandit methodology. See the work of (Brochu et al., 2010) for an extensive tutorial. An explicit link between Bayesian optimization and bandit methods has also been drawn in the work of (Srinivas et al., 2010), wherein the authors develop a UCB approach using Gaussian processes to model the reward distributions. The authors also provide bounds on the cumulative regret of this method.

Up to this point, however, all the Bayesian methods we have mentioned focus on the problem of cumulative rewards. A natural question to ask then, is how to incorporate Bayesian methodology into the problem of best arm identification—most of the literature on best arm identification approaches this problem from a frequentist perspective. In the setting of Bayesian optimization, the work of (Hennig & Schuler, 2012) formulates the global optimization problem and describes several techniques and approximations to make the inference problem tractable. In this work, however, the authors mainly focus on algorithmic contributions and do not provide any theoretical performance bounds. Our work features the first, to our knowledge, Bayesian approach targeting the problem of best arm identification with provable performance bounds with stochastic rewards. Alternatively, (de Freitas et al., 2012) show similar performance bounds to ours in the setting of Bayesian optimization with deterministic rewards.

First, in Section 2 we formally introduce the problem of best arm identification. Then, in Section 3, we introduce the method that we will use throughout the rest of this work, which builds on the gap-based exploration approach introduced in (Gabillon et al., 2011; 2012). We provide a key generalization to the approach of Gabillon et al. which is crucial in deriving bounds for Bayesian models. Furthermore, our theorem subsumes the regret bounds for the frequentist methods based on Hoeffding and Empirical Bernstein bounds. Still in Section 3, we state Theorem 1 which bounds the regret in this general problem, and discuss its implications. The proof of this theorem, given in the appendix, follows that of the earlier work incorporating our generalization. In Section 4 we introduce a Bayesian framework for gap-based exploration, which we call BayesGap. In particular, we consider the case of linear-Gaussian arms and use the result of Section 3 to give a bound on this method's performance. To demonstrate the generality of the proof of Theorem 1 we include a bound for Bernoulli bandits in the appendix (see the supplementary material). Finally, in Section 5 we consider the empirical performance of BayesGap as compared to a number of different approaches from the literature on independent Gaussian arms, correlated Gaussian arms, and finally on a real-world sensor network optimization task.

2. Problem formulation

Consider a multi-armed bandit problem with a collection of independent arms $\mathcal{A} = \{1, \ldots, K\}$ such that the immediate reward of pulling arm $k \in \mathcal{A}$ is characterized by a distribution ν_k with mean μ_k. Note that the assumption of independence does *not* mean that the means of each arm cannot share some underlying structure—only that the act of pulling arm k does not affect the future rewards of pulling any other arm. This distinction will be relevant in Section 4.1.

Next, we will let

$$\Delta_k = |\max_{i \neq k} \mu_i - \mu_k| \tag{1}$$

denote the difference between the kth arm and the best alternative arm. For the optimal arm this coincides with a measure of how optimal that arm is, whereas for all other arms it is a measure of their sub-optimality. Finally, we will also let μ^* denote the best arm where k^* is its corresponding index.

The problem of identifying the best arm in a multi-armed bandit process can now be introduced as a sequential decision problem. At each round t the decision maker will select or "pull" an arm $a_t \in \mathcal{A}$ and observe an independent sample y_t drawn from the corresponding distribution ν_{a_t}. At the beginning of each round t, the decision maker must decide which arm to select based only on previous interactions, which we will denote with the tuple $(a_{1:t-1}, y_{1:t-1})$. For any arm k we can also introduce the immediate, expected regret of selecting that arm as

$$R_k = \mu^* - \mu_k, \tag{2}$$

i.e. the difference between the expected reward of selecting arm k versus the reward of selecting the best arm.

In standard bandit problems the goal is generally to minimize the cumulative sum of immediate regrets incurred by the arm selection process. Instead, in this work we consider the *pure exploration* setting which divides the sampling process into two phases: exploration and testing. The exploration phase consists of T rounds wherein a decision maker interacts with the bandit process by sampling arms. After these rounds, the decision maker must make a single arm recommendation $\Omega(T) \in \mathcal{A}$. The performance of the decision maker is then judged only on the performance of this recommendation strategy. The expected performance of this single recommendation is known as the *simple regret*, and we can write this quantity as $R_{\Omega(T)}$. Given an $\epsilon > 0$ we can then define the *probability of error* as the probability that $R_{\Omega(T)} > \epsilon$.

Somewhat surprisingly, (Bubeck et al., 2009) shows that any arm selection strategy that attains the optimal, logarithmic cumulative regret, i.e. of order $\log(t)$ obtains non-cumulative regret of order $t^{-\gamma}$ for some γ. As a result, the only way to obtain exponentially vanishing probability of error is to abandon the optimal cumulative rate and explore more aggressively (see Bubeck et al., 2009; Audibert et al., 2010). The Bayesian gap-based approach, as well as the more general gap-based approach we discuss, are both able to obtain exponentially vanishing rate, whereas standard cumulative regret methods *provably do not*. The experimental results on this behavior are somewhat more subtle, however, and we will return to this in Section 5.

3. General gap-based exploration

At the beginning of round t we will assume that the decision maker is equipped with upper and lower bounds $U_k(t)$ and $L_k(t)$ on the mean reward for the kth arm. For the time being we will make no assumption on these bounds other than that with probability at least $1 - \delta$ they bound the mean μ_k, i.e.

$$\Pr(L_k(t) \leq \mu_k \leq U_k(t)) \geq 1 - \delta. \tag{3}$$

We will also introduce the uncertainty diameter $s_k(t) = U_k(t) - L_k(t)$ associated with each arm k. Given bounds on the mean reward for each arm, we can then introduce the gap quantity

$$B_k(t) = \max_{i \neq k} U_i(t) - L_k(t), \tag{4}$$

which we can easily see involves a comparison between the lower bound of arm k and the highest upper bound among all alternative arms. Ultimately the arm selection strategy will be based on this index. However, rather than directly finding the arm minimizing this gap, we will consider the two arms whose upper and lower bounds define this minimizer, namely

$$J(t) = \operatorname*{arg\,min}_{k \in \mathcal{A}} B_k(t) \quad \text{and} \quad j(t) = \operatorname*{arg\,max}_{k \neq J(t)} U_k(t).$$

Algorithm 1

General gap-based exploration algorithm.

```
init: select ea. arm once to obtain $(a_{1:K}, y_{1:K})$
for $t = K + 1, \ldots$ do
    compute $L_k(t), U_k(t)$, and $B_k(t)$
    $J(t) = \arg\min_{k \in A} B_k(t)$
    $j(t) = \arg\min_{k \neq J(t)} U_k(t)$
    select arm $a_t = \arg\max_{k \in \{j(t), J(t)\}} s_k(t)$
    observe $y_t \sim \nu_{a_t}(\cdot)$
    break if termination condition is true
end for
return $\Omega(t)$
```

We can then select action a_t such that

$$a_t = \operatorname*{arg\,max}_{k \in \{j(t), J(t)\}} s_k(t), \tag{5}$$

i.e. between these two arms select the arm with the greatest uncertainty.

In Algorithm 1 we show a general algorithm for the problem of gap-based exploration, however note that we have also not defined the termination condition or the recommendation strategy $\Omega(t)$. In this work we will consider the case of a fixed arm-selection budget, where the decision-maker must make exactly T arm queries. We note, however, that it is possible to extend these strategies to a setting where the decision maker can take as many actions as necessary to reach some desired certainty. The ability to handle both bounded horizon and bounded uncertainty is the main driver of the unified approach of (Gabillon et al., 2012); in this work we do not address the task of bounded confidence purely for reasons of simplicity.

In the budgeted horizon setting the termination condition is simple: once $t = T$ and the time-horizon is reached we must break out of the loop. We will then define the recommendation strategy as

$$\Omega(T) = J\big(\operatorname*{arg\,min}_{t \leq T} B_{J(t)}(t)\big). \tag{6}$$

Here we can see that this corresponds to finding the proposal arm $J(t)$ which corresponds to the minimum over all bounds, over all times $t \leq T$. The reason behind this particular choice is subtle and will become clear in the proof of Theorem 1 in the appendix.

We will first define $N_k(t)$ as the number of times arm k has been pulled after t rounds. Theorem 1 is most powerful when the behaviour of $s_k(t)$ is known as a function of $N_k(t - 1)$. Since this exact relationship can be hard to compute, we propose to use an upper bound on $s_k(t)$ instead: the tighter the bound, the better the result. We will let $g_k : \mathbb{N} \to \mathbb{R}^+$ be a strictly monotonically decreasing function such that

$$s_k(t) \leq g_k(N_k(t-1)). \tag{7}$$

One important feature of g_k that we exploit in Theorem 1 is that, as a result of being monotonically decreasing it must be injective and therefore reversible. In a slight abuse of notation we will let g_k^{-1} denote the left inverse of g_k. Often, for unstructured and independent arms we can utilize the same bound $g_k = g$ for each arm. This generalization does, however, allow us to incorporate model information on an arm-by-arm basis, as we will see in the next section.

In order to properly explore the bandit problem we will need to control how many times each arm k is pulled based on how difficult it is to determine whether that arm is optimal with accuracy ϵ. We will do so by introducing an arm-dependent hardness quantity

$$H_{k\epsilon} = \max(\tfrac{1}{2}(\Delta_k + \epsilon), \epsilon). \tag{8}$$

and $H_\epsilon = \sum_k H_{k\epsilon}^{-2}$ as a problem-dependent hardness parameter associated with the bandit problem as a whole. When it is not specified, the term hardness refers to H_ϵ. We will see this quantity reappear in a number of problem-specific bounds shown in the next section.

Theorem 1. *Consider a bandit problem with horizon T and K arms. Let $U_k(t)$ and $L_k(t)$ be upper and lower bounds that hold for all times $t \leq T$ and all arms $k \leq K$ with probability $1 - \delta$. Finally, let g_k be a monotonically decreasing bound on the confidence diameter for arm k, as defined in* (7)*, such that $\sum_k g_k^{-1}(H_{k\epsilon}) \leq T - K$. We can then bound the simple regret as*

$$\Pr(R_{\Omega(T)} \leq \epsilon) \geq 1 - KT\delta. \tag{9}$$

The result of Theorem 1 is general enough to accommodate a large class of uncertainty models, whether frequentist or Bayesian. In addition, the theorem reduces the problem of proving a regret bound to that of checking a few properties of the uncertainty model. For example, by using Hoeffding or Bernstein bounds to define the confidence intervals we recover the bounds of (Gabillon et al., 2012). In the following section, we will apply this theorem to Bayesian Gaussian bandits in order to obtain a bound on the performance in terms of simple regret. We further demonstrate the flexibility of this theorem in the appendix where the same technique is applied to Bernoulli bandits.

4. Bayesian gap-based exploration

In this section, we will consider bandit problems wherein the distribution of rewards for each arm k is assumed to depend on unknown parameters $\theta_k \in \Theta$. We will write the density of each arm as ν_{θ_k}. When considering the bandit problem from a Bayesian perspective, we will assume a prior density $\theta_k \sim \pi_k^0(\cdot)$ from which the parameters for each arm are drawn. After $t-1$ rounds, let $T_k(t-1) = \{n < t : a_n = k\}$ be the the subset of past time indices such that arm k was selected. Given these indices, the posterior for the parameters of arm k can be written as

$$\pi_k^t(\theta_k) \propto \pi_k^0(\theta_k) \prod_{n \in T_k(t-1)} \nu_{\theta_k}(y_n). \tag{10}$$

From this we can also see that only if arm k is selected at time $t-1$ does the posterior need to be updated, as otherwise for all $k \neq a_{t-1}$ we have $\pi_k^t = \pi_k^{t-1}$.

We are, however, only partially interested in the posterior distribution of the parameters θ_k. Instead, we are primarily concerned with the posterior distribution of the mean reward for each arm

$$\mu_k = \mathbb{E}[Y|\theta_k] = \int y \, \nu_{\theta_k}(y) \, dy. \tag{11}$$

Again, we note that we do not know the true value of θ_k, instead we only have access to the posterior distribution over this variable at time t. As a result we only have a distribution over μ_k, induced by the distribution over the parameters, which at time t we can write as $\rho_k^t(\mu_k)$.

In the subsections that follow we will use this posterior distribution to define upper and lower confidence bounds that both hold with high probability and give rise to a bound on the confidence diameter g_k of the desired form. As a result, we can derive high-probability bounds on the simple regret which can easily take into account the structure of the problem.

4.1. Gaussian arms

Consider K arms, such that each arm k is associated with a known vector $u_k \in \mathbb{R}^d$ and where the rewards for pulling arm k are normally distributed

$$\nu_k(y) = \mathcal{N}(y; u_k^T\theta, \sigma^2) \tag{12}$$

with known noise σ^2 and unknown $\theta \in \mathbb{R}^d$. Note here that the rewards for each arm are independent conditional on θ, but marginally dependent when this parameter is unknown. In particularly the level of their dependence depends on the structure of the vectors

u_k. By placing a prior $\theta \sim \mathcal{N}(0, \eta^2 I)$ over the entire parameter vector, however, we can still compute a posterior distribution over this unknown quantity.

Let $X_t = [u_{a_1} \dots u_{a_{t-1}}]^T$ denote the design matrix and $Y_t = [y_1 \dots y_{t-1}]^T$ the vector of observations at the beginning of round t. We can then write the posterior at time t as $\pi^t(\theta) = \mathcal{N}(\theta; \hat{\theta}_t, \hat{\Sigma}_t)$, where

$$\hat{\Sigma}_t^{-1} = \sigma^{-2} X_t^T X_t + \eta^{-2} I, \tag{13}$$

$$\hat{\theta}_t = \sigma^{-2} \hat{\Sigma}_t X_t^T Y_t. \tag{14}$$

From this formulation we can easily obtain that the expected reward associated with arm k is marginally normal $\rho_k^t(\mu_k) = \mathcal{N}(\mu_k; \hat{\mu}_k(t), \hat{\sigma}_k^2(t))$ with mean $\hat{\mu}_k(t) = u_k^T \hat{\theta}_t$ and variance $\hat{\sigma}_k^2(t) = u_k^T \hat{\Sigma}_t u_k$. Note also that the predictive distribution over rewards associated with the kth arm is normal as well, with mean $\hat{\mu}_k(t)$ and variance $\hat{\sigma}_k^2(t) + \sigma^2$.

Finally, based on this posterior, we will introduce upper and lower bounds given by

$$L_k(t) = \hat{\mu}_k(t) - \beta\, \hat{\sigma}_k(t), \tag{15}$$

$$U_k(t) = \hat{\mu}_k(t) + \beta\, \hat{\sigma}_k(t), \tag{16}$$

from which we can then claim the following:

Corollary 1. *Consider a K-armed Gaussian bandit problem with horizon T and let $U_k(t)$ and $L_k(t)$ be defined as above. Let $\kappa = \sum_k \|u_k\|^{-2}$. Then for $\epsilon > 0$ and*

$$\beta^2 = \big((T-K)/\sigma^2 + \kappa/\eta^2\big)/(4H_\epsilon),$$

the algorithm attains simple regret satisfying

$$\Pr(R_{\Omega(T)} \le \epsilon) \ge 1 - KTe^{-\beta^2/2}$$

Proof. First, to simplify notation let $\lambda = \sigma/\eta$. Now, using the definition of the posterior variance for arm k as $\hat{\sigma}_k^2(t)$, we can write the confidence diameter as

$$\begin{aligned} s_k(t) &= 2\beta\sqrt{u_k^T \hat{\Sigma}_t u_k} \\ &= 2\beta\sqrt{\sigma^2 u_k^T \big(\textstyle\sum_i N_i(t-1)\, u_i u_i^T + \lambda^2 I\big)^{-1} u_k} \\ &\le 2\beta\sqrt{\sigma^2 u_k^T \big(N_k(t-1)\, u_k u_k^T + \lambda^2 I\big)^{-1} u_k}. \end{aligned}$$

In the second equality we decomposed the Gram matrix $X_t^T X_t$ in terms of a sum of outer products over the fixed vectors u_i. In the final inequality we noted that by removing samples we can only increase the variance term, i.e. here we have essentially replaced $N_i(t-1)$ with 0 for $i \ne k$. We will let the result of this final inequality define an arm-dependent bound g_k. Letting $A = \lambda^2/N$ we can simplify this quantity using the Sherman-Morrison formula as

$$\begin{aligned} g_k(N) &= 2\beta\sqrt{(\sigma^2/N) u_k^T \big(u_k u_k^T + AI\big)^{-1} u_k} \\ &= 2\beta\sqrt{\frac{\sigma^2}{N}\frac{\|u_k\|^2}{A}\Big(1 - \frac{\|u_k\|^2/A}{1 + \|u_k\|^2/A}\Big)} \\ &= 2\beta\sqrt{\frac{\sigma^2\|u_k\|^2}{\lambda^2 + N\|u_k\|^2}}, \end{aligned}$$

which is monotonically decreasing in N. The inverse of this function can easily be solved for as

$$g_k^{-1}(s) = \frac{4(\beta\sigma)^2}{s^2} - \frac{\lambda^2}{\|u_k\|^2}.$$

By setting $\sum_k g_k^{-1}(H_{k\epsilon}) = T - K$ and solving for β we then obtain the definition of this term given in the statement of the corollary. Finally, by reference to Lemma D1 we can see that for each k and t, the upper and lower bounds must hold with probability $1 - e^{-\beta/2}$. These last two statements satisfy the assumptions of Theorem 1, thus concluding our proof. □

5. Experiments

We can now turn to the problem of empirically comparing our proposed algorithm, BayesGap, to several other approaches advocated in the literature for the linear-Gaussian model introduced in Section 4.1. In particular we will consider the following approaches:

1. *UCBE*: A *highly exploring* variant of the classical UCB policy of (Auer et al., 2002), introduced by (Audibert et al., 2010). This approach replaces the $\log(t)$ exploration term of UCB with a constant of order $\log(T)$ for known horizon T. This encourages the algorithm to explore much more aggressively.

2. *UGap*: The bounded-horizon[1] gap-based exploration approach introduced in (Gabillon et al., 2012) and which Section 3 is based on. As mentioned in that section, the algorithm is based on using confidence bounds derived from a Hoeffding bound[2] around the mean.

3. *BayesUCB*: An adaption of UCB to the Bayesian setting wherein the upper confidence bound is

[1] Technically this is UGapEb, denoting bounded horizon, but as we do not consider the fixed-confidence variant in this paper we simplify the acronym.

[2] The authors also introduced a variation, UGapV, which uses tighter Bernstein bounds. However, in this paper we restrict our comparison to UGap. We note also that for bounded horizon problems in this earlier work, UGap and UGapV obtained similar results.

given by an upper quantile on the posterior mean (see Kaufmann et al., 2012a).

4. *Thompson sampling*: A randomized, Bayesian index strategy where the probability of selecting the kth arm is given by a single-sample Monte Carlo approximation to the posterior probability that arm k is the best arm. See also (Chapelle & Li, 2012) for an empirical study, and (Kaufmann et al., 2012b) for a theoretical analysis.

Among these algorithms, (1–2) attack the pure exploration problem, whereas (3–4) are optimal for cumulative regret problems. Algorithms (3–4) are also Bayesian approaches. For the cumulative regret approaches we used as recommendation strategy the index of the best posterior mean after T rounds. Note also that we did not compare against the classical UCB algorithm due to the fact that its bounds are sub-optimal compared to those of BayesUCB, a fact that is also borne out empirically (see the above citations).

We also note that in the linear-Gaussian setting we analyze here, the approach of BayesUCB—as pointed out by (Kaufmann et al., 2012a)—coincides with the linear optimization approach of (Dani et al., 2008) for a particular choice of uninformative prior. As a result, our method would also be able to take advantage of this prior structure by replacing the simpler $\mathcal{N}(0, \eta^2 I)$ prior to account for unknown variance.

Finally, although (3–4) yield competitive results in the following experiments, the fact that these methods achieve the optimal logarithmic cumulative regret bound (Kaufmann et al., 2012a) implies that they are provably sub-optimal in the simple regret setting (Bubeck et al., 2009). In contrast, BayesGap has provable, exponentially vanishing simple regret (see Corollary 1), along with UGap and UCBE.

For the remainder of the section, unless otherwise specified, the Bayesian methods are given a Gaussian prior over θ with mean zero and variance $\eta^2 = 1$. The observation model in the synthetic experiments is also given to the methods and is set to a Gaussian with variance $\sigma^2 = 0.25$.

For the simple regret approaches, we also give each algorithm the hardness estimate H_ϵ. In practice one would not know this parameter, and instead would have to estimate it in an adaptive fashion as done in (Audibert et al., 2010; Gabillon et al., 2011). We did not do this for simplicity and also for a more direct comparison with the results of the most closely related algorithm, that of (Gabillon et al., 2012). Finally, as is often the case in the literature (see the above citations), we found that tuning the exploration parameter of each of these approaches subtly improved their behavior. In the case of BayesGap this amounts to using a parameter $\overline{\beta} = c\beta$ for some constant $c > 0$, and similarly for the other two methods. We should also note that while UCBE and UGap assume bounded rewards, the use of an exploration multiplier c for both algorithms helps to correct for the fact that our rewards are actually unbounded. One could consider different bounding techniques in order to extend these models to unbounded rewards, but we should point out that moving into the Bayesian setting is one particular way to take into account rewards of this form.

The next four sections are each dedicated to a separate experiment. In Section 5.1, we evaluate the sensitivity of the tested methods to their hyper-parameters. In Section 5.2, we fix the hardness H_ϵ and evaluate the methods for various horizons. The next two sections introduce correlation via Gaussian process (GP) techniques. In Section 5.3, we present an experiment where the means μ_k are set by sampling a function at discrete points, a function which was itself sampled from a synthetic GP. In Section 5.4, we obtain μ_k similarly from a GP that was fit to real data.

5.1. Sensitivity Analysis

We first present experiments on a synthetic set of Gaussian arms in order to analyze the sensitivity of BayesGap, UCBE, and UGap, to their hyper-parameters. We considered $K = 20$ independent Gaussian arms, and as noted earlier we set the prior and likelihood (observation noise) variances to $\sigma^2 = 0.25$ and $\eta^2 = 1$ for BayesGap. We ran multiple experiments by varying the hardness of the problem with $H_\epsilon \in \{40, 80, 160, 320\}$; note that the choice of H_ϵ corresponds to four multiples of K/σ^2. We used this choice of H_ϵ to set the arm means such that each run includes an optimal arm of $\mu^* = \eta/4 + \sqrt{K/H_\epsilon}$ and all other arms have mean $\mu_k = \eta/4$. We also varied the time horizon, using $T \in \{25, 40, 80, 160\}$. Finally, for each (T, H_ϵ) pair we studied the performance of the three algorithms using a regular, logarithmic grid for the exploration parameter c. A few typical examples are reported in Figure 1.

We gained a number of insights from these experiments. First, BayesGap's optimal performance for fixed T and H_ϵ is comparable, if not better, than UGap and UCBE. In fact, for short horizons on "easy" problems, BayesGap outperforms them significantly. However this is often a regime of general poor performance for all methods as is illustrated in the two lower-left plots of Figure 1. Second, for low values of c, BayesGap is more sensitive to its parameter than its

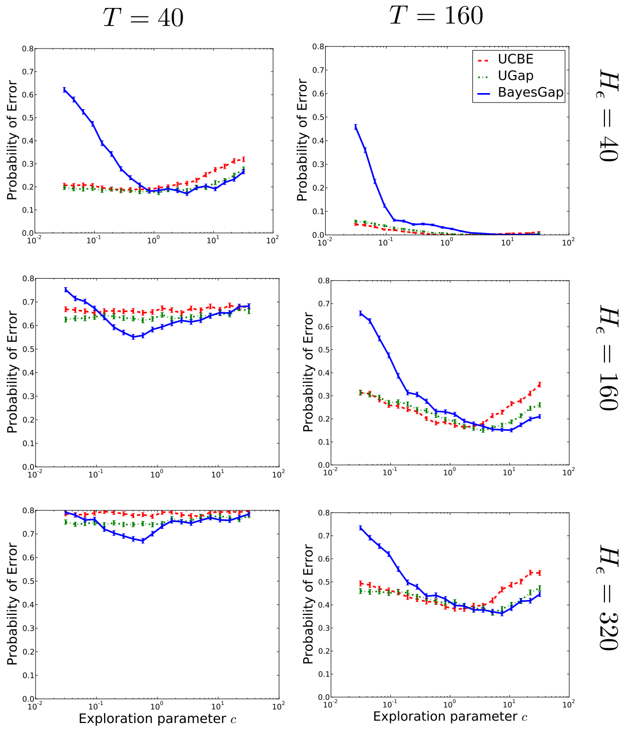

Figure 1. Sensitivity of UCBE, UGap, and BayesGap to the tuning parameter c for varying horizons T and hardness quantities H_ϵ.

competition, while for large enough c, the behaviour is largely similar as can be seen in the left column of Figure 1. Finally, BayesGap's optimal c value seems to drift towards larger values as we look at larger horizons. We found in most of our experiments that a value in the range $[1, 8]$ was appropriate. Only when we attempted to drastically increase the hardness of our problem did larger values perform somewhat better.

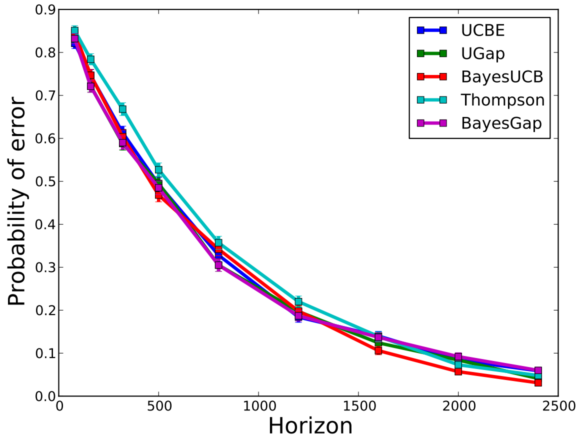

Figure 2. All the methods evaluated at different time horizons. The exploration parameter for UCBE, UGap, and BayesGap is set to $c = 8$. The probability of error is estimated using Monte Carlo with $N = 1000$ samples. Error bars show ±1 standard error.

5.2. Multiple Fixed Horizons

In this experiment, we consider the bandit problem from the previous subsection with $H_\epsilon = 1280$ and we fix $c = 8$ for all methods. This value of the hyperparameter produced good results for all three methods in the previous experiment. We also consider a harder problem here than in the previous section to show the behaviour of these algorithms for long horizons. We then measure the performance of the algorithms with this value for a large range of fixed horizons. It is *important to note* that this is not showing the evolution in time of the performance of a fixed algorithm. Rather each point in Figure 2 is an average over multiple independent runs of the algorithm for a given horizon T. (Indeed, recall that a different T corresponds to a different exploration trade-off and therefore a new instance of the algorithm.) Since c and H_ϵ are held fixed and T is being varied, each point in Figure 2 is an entirely different algorithm instance. The main point to take away from this figure is that BayesGap performs as well as its competitors in the independent arms setting. It is perhaps somewhat surprising how well the cumulative regret methods do. This, however, somewhat matches previous observations from the literature in that the comparatively "poor" performance of these methods may only appear for very hard instances or in the asymptotic regime; see (Bubeck et al., 2009, Figure 4 and discussion).

Notice that for large horizons, BayesUCB and Thompson seem to slightly outperform BayesGap. Given our sensitivity analysis results, this is not surprising, since for large horizons, the optimal c value for BayesGap exhibits a drift towards larger values. Moreover, Corollary 1 above and Theorem 1 from (Gabillon et al., 2012) guarantee that the probability of error for BayesGap and UGap vanishes exponentially fast in the limit, respectively. Conversely, BayesUCB and Thompson are provably sub-optimal due to the previously discussed result of (Bubeck et al., 2009).

5.3. Application to a Synthetic Optimization Problem of Varying Correlation

In this experiment, we study the effect of correlation among the arms. If the structure of the correlation is known, then each arm pull possibly provides information about more than one arm. If this information is used judiciously, greater performance can be attained.

In order to create a problem with known, measurable correlation structure, we use a Gaussian process (see Rasmussen & Williams, 2005) evaluated at a discrete number of points using a squared-exponential kernel $k(x, x')$ with length-scale ℓ^2. This allows us to control the correlation between arms by adjusting ℓ, where larger ℓ corresponds to larger correlations. More precisely, given points $\{x_i\}$ we can construct a matrix G such that $G_{ij} = k(x_i, x_j)$. The matrix G is usually denoted K in the GP literature, however our notation departs from this standard order to distinguish from the number of arms K. Translating this into a Bayesian linear model is then simply a matter of taking the SVD, $G = VDV^T$ for diagonal D and unitary V, and constructing the design matrix $U = VD^{\frac{1}{2}}$. We can then think of each arm k as corresponding to a particular element in the set $\{x_i\}$. Then for $\theta \sim \mathcal{N}(0, \eta^2)$ we have that $\mu_k = u_k^T\theta$ is a sample from a Gaussian process with the given kernel and signal variance η^2. Essentially we are performing Bayesian optimization using a GP prior at a discrete number of points; see (Srinivas et al., 2010). In fact, in our setting BayesUCB corresponds to GP-UCB with a slightly different arm-selection mechanism.

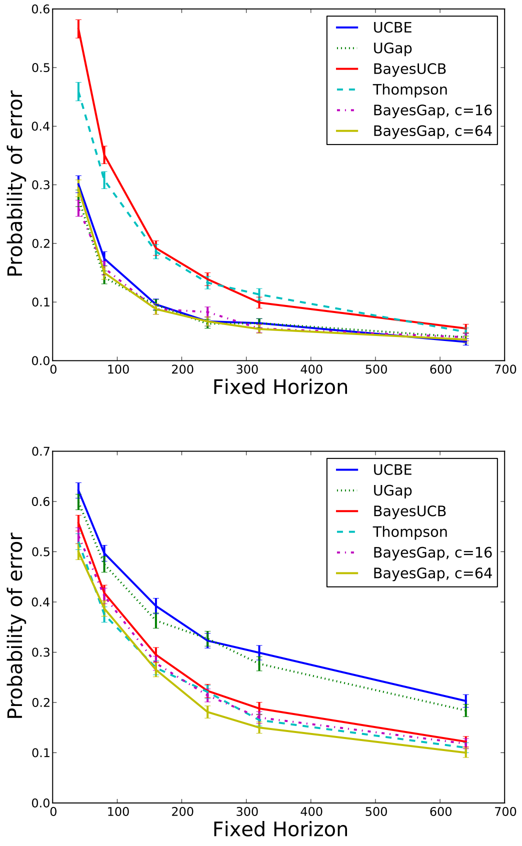

Figure 3. Probability of error for multiple fixed horizons with **(top)** $\ell = 0.25$ and **(bottom)** $\ell = 3.0$, corresponding to low and high correlation, respectively. The probability of error is estimated using Monte Carlo $N = 1000$ samples. Error bars show ±1 standard error.

A side effect of this type of correlation structure is that the best arm is now necessarily surrounded by very good arms. This gap between them decreases with increasing ℓ which in turn makes the problem harder. Given this correlation structure we next considered 30 arms, corresponding to 30 points on a simple 1d-grid. For two length scales $\ell = 0.25$ and $\ell = 3.0$ we then performed $N = 1000$ runs of each bandit approach, each averaged using individual samples θ from the prior. A plot similar to the one in the previous section is shown in Figure 3 for both of these length scales. In order to avoid diverging problem hardnesses, here we set $\epsilon = 0.1$ to upper bound the hardness of the problem by $K/\epsilon^2 = 3000$. This results in a value of H_ϵ which is higher than any problem we have considered so far, and at this level of difficulty we found that larger values of c performed better, as was suggested by our discussion at the end of Section 5.1.

The results reported in Figure 3 show that, as expected, for low correlation (top), BayesGap performs as well as the frequentist methods, while the other Bayesian methods lack exploration. Meanwhile, for high correlation (bottom), since the Bayesian methods incorporate the correlation structure in their posterior, the less exploratory BayesUCB and Thompson now achieve good performance, whereas the frequentist methods are out-performed because they are agnostic to the correlation structure. Note that for both length scales, BayesGap is one of the best performing methods.

5.4. Application to Real Data

Finally, we take the same Gaussian process based approach from the previous set of experiments and apply it to real data. In particular we utilized data taken from traffic speed sensors deployed along highway I-880 South in California. This data, was also used in (Srinivas et al., 2010). Traffic speed was collected for all working days between 6AM and 11AM for one month using 357 sensors. The goal is then to identify the single location with the highest expected speed, i.e. the least congested.

Rather than specifying a kernel over the space of traffic sensor locations we follow the approach of (Srinivas et al., 2010) and construct the matrix by treating two-thirds of the data as historical data, letting the kernel matrix G then be given by the empirical covariance of

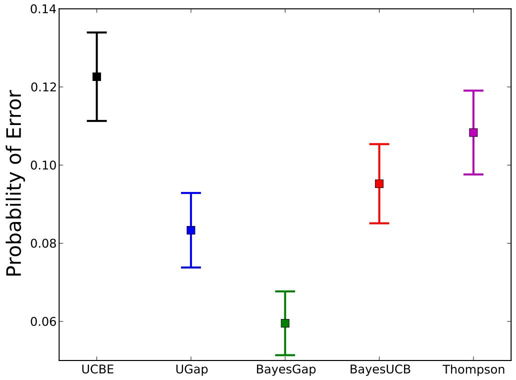

Figure 4. Probability of error on the optimization domain of traffic speed sensors. For this real data set, BayesGap provides considerable improvements over the Bayesian cumulative regret alternatives and the frequentist simple regret counterparts.

this dataset. We then applied the same transformation as above to construct the design matrix U for Bayesian linear regression. We also took the averaged variance of each individual sensor (i.e. the signal variance) and set the noise variance σ^2 of our process equal to 5% of this value (4.78). Given this historical data we also selected, somewhat arbitrarily, a broad prior of $\eta = 20$ (note that this data *does* correspond to traffic speeds in California).

Finally, to evaluate this experiment we performed a single run of each bandit method using each of the remaining sensor signals as the mean vector μ with the given noise σ and design matrix U. The results shown in Figure 4 are the probability of error for each method using a time horizon of $T = 400$. Here we used as c for each of the relevant methods the value suggested from our earlier sensitivity experiments; for BayesGap this corresponded to $c = 8$.

We can clearly see here the value of BayesGap in that it combines the best of both the pure exploration properties of UGap with the ability to take advantages of correlations via the Bayesian prior/posterior.

6. Conclusion

In this work we presented, to our knowledge, the first Bayesian approach to the problem of best arm identification with provable, exponentially vanishing, high probability regret bounds. In order to do so we built upon the earlier, gap-based approach of (Gabillon et al., 2011; 2012). We provided a key generalization to this earlier approach which accomodates Bayesian uncertainty models, non-symmetric confidence diameters, and simplifies the process of proving model-specific simple regret bounds. We then applied this approach to the problem of linear-Gaussian bandits, and in the appendix we provide bounds for Bernoulli bandits.

Our experiments show that our proposed method is competitive with existing approaches both in the presense and absense of correlated arm structure—and always among the best exploration strategies. Further, where other state-of-the-art bandit techniques for linear bandits have provably sub-optimal simple regret bounds, our approach is able to obtain exponentially vanishing regret. Finally, we showed that in a realistic optimization example our method can significantly outperform other existing techniques.

References

J.-Y. Audibert, S. Bubeck, and R. Munos. Best arm identification in multi-armed bandits. In *Proceedings of the Conference on Learning Theory*, 2010.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

D. Berry and B. Fristedt. *Bandit Problems: Sequential Allocation of Experiments.* Chapman & Hall, 1985.

E. Brochu, V. Cora, and N. de Freitas. A tutorial on Bayesian optimization of expensive cost functions with application to active user modeling and hierarchical reinforcement learning. eprint arXiv:1012.2599, arXiv, 2010.

S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in multi-armed bandits problems. In *the International Conference on Algorithmic Learning Theory*, 2009.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games.* Cambridge University Press, New York, 2006.

O. Chapelle and L. Li. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems*, 2012.

V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic Linear Optimization under Bandit Feedback. In *Proceedings of the Conference on Learning Theory*, pp. 355–366, 2008.

N. de Freitas, A. J. Smola, and M. Zoghi. Exponential Regret Bounds for Gaussian Process Bandits with

Deterministic Observations. In *International Conference on Machine Learning*, 2012.

V. Gabillon, M. Ghavamzadeh, A. Lazaric, and S. Bubeck. Multi-bandit best arm identification. In *Advances in Neural Information Processing Systems*, 2011.

V. Gabillon, M. Ghavamzadeh, and A. Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems*, 2012.

J. Gittins, K. Glazebrook, and R. Weber. *Multi-armed bandit allocation indices*. Wiley, 2011.

P. Hennig and C. J. Schuler. Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 13:1809–1837, 2012.

E. Kaufmann, O. Cappé, and A. Garivier. On Bayesian upper confidence bounds for bandit problems. In *Artificial Intelligence and Statistics*, 2012a.

E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: an asymptotically optimal finite-time analysis. In *the International Conference on Algorithmic Learning Theory*, 2012b.

R. Kohavi, R. Longbotham, D. Sommerfield, and R. Henne. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery*, 18:140–181, 2009.

T. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6, 1985.

C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 55, 1952.

D. Russo and B. Van Roy. Learning to optimize via posterior sampling. eprint arXiv:1301.2609, arXiv, 2013.

S. Scott. A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6), 2010.

N. Srinivas, A. Krause, S. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning*, 2010.

W. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

A. Proof of Theorem 1

Note that the proof of this section and the lemmas of the next section follow from the proofs of (Gabillon et al., 2012), but generalized to accomodate more general functions g and nonsymmetric confidence diameters s_k.

Proof. We will first define the event $\mathcal{E}$ such that on this event every mean is bounded by its associated bounds for all times t. More precisely we can write this as

$$\mathcal{E} = \{\forall k \leq K, \forall t \leq T, L_k(t) \leq \mu_k \leq U_k(t)\}.$$

By definition, these bounds are given such that the probability of deviating from a single bound is δ. Using a union bound we can then bound the probability of remaining within all bounds as $\Pr(\mathcal{E}) \geq 1 - KT\delta$.

We will next condition on the event $\mathcal{E}$ and assume regret of the form $R_{\Omega(T)} > \epsilon$ in order to reach a contradiction. Upon reaching said contradiction we can then see that the simple regret must be bounded by ϵ with probability given by the probability of event $\mathcal{E}$, as stated above. As a result we need only show that a contradiction occurs.

We will now define $\tau = \arg\min_{t \leq T} B_{J(t)}(t)$ as the time at which the recommended arm attains the minimum bound, i.e. $\Omega(T) = J(\tau)$ as defined in (6). Let $t_k \leq T$ be the last time at which arm k is pulled. Note that each arm must be pulled at least once due to the initialization phase. We can then show the following sequence of inequalities:

$$\begin{aligned} \min(0, s_k(t_k) - \Delta_k) + s_k(t_k) &\geq B_{J(t_k)}(t_k) && \text{(a)} \\ &\geq B_{\Omega(T)}(\tau) && \text{(b)} \\ &\geq r_{\Omega(T)} && \text{(c)} \\ &> \epsilon. && \text{(d)} \end{aligned}$$

Of these inequalities, (a) holds by Lemma B3, (c) holds by Lemma B1, and (d) holds by our assumption on the simple regret. The inequality (b) holds due to the definition $\Omega(T)$ and time τ. Note, that we can also write the preceding inequality as two cases

$$\begin{aligned} s_k(t_k) > 2s_k(t_k) - \Delta_k > \epsilon, &\quad \text{if } \Delta_k > s_k(t_k); \\ 2s_k(t_k) - \Delta_k \geq s_k(t_k) > \epsilon, &\quad \text{if } \Delta_k \leq s_k(t_k) \end{aligned}$$

This leads to the following bound on the confidence diameter,

$$s_k(t_k) > \max(\tfrac{1}{2}(\Delta_k + \epsilon), \epsilon) = H_{k\epsilon}$$

which can be obtained by a simple manipulation of the above equations. More precisely we can notice that in each case, $s_k(t_k)$ upper bounds both ϵ and $\frac{1}{2}(\Delta_k + \epsilon)$, and thus it obviously bounds their maximum.

Now, for any arm k we can consider the final number of arm pulls, which we can write as

$$\begin{aligned} N_k(T) = N_k(t_k - 1) + 1 &\leq g^{-1}(s_k(t_k)) + 1 \\ &< g^{-1}(H_{k\epsilon}) + 1. \end{aligned}$$

This holds due to the definition of g as a monotonic decreasing function, and the fact that we pull each arm at least once during the initialization stage. Finally, by summing both sides with respect to k we can see that $\sum_k g^{-1}(H_{k\epsilon}) + K > T$, which contradicts our definition of g in the Theorem statement. □

B. Lemmas

In order to simplify notation in this section, we will first introduce $B(t) = \min_k B_k(t)$ as the minimizer over all gap indices for any time t. We will also note that this term can be rewritten as

$$B(t) = B_{J(t)}(t) = U_{j(t)}(t) - L_{J(t)}(t),$$

which holds due to the definitions of $j(t)$ and $J(t)$.

Lemma B1. *For any sub-optimal arm $k \neq k^*$, any time $t \in \{1, \dots, T\}$, and on event $\mathcal{E}$, the immediate regret of pulling that arm is upper bounded by the index quantity, i.e. $B_k(t) \geq r_k$.*

Proof. We can start from the definition of the bound and expand this term as

$$\begin{aligned} B_k(t) &= \max_{i \neq k} U_i(t) - L_k(t) \\ &\geq \max_{i \neq k} \mu_i - \mu_k = \mu^* - \mu_k = r_k. \end{aligned}$$

The first inequality holds due to the assumption of event $\mathcal{E}$, whereas the following equality holds since we are only considering sub-optimal arms, for which the best alternative arm is obviously the optimal arm. □

Lemma B2. *For any time t let $k = a_t$ be the arm pulled, for which the following statements hold:*

$$\begin{aligned} &\textit{if } k = j(t), \textit{ then } L_{j(t)}(t) \leq L_{J(t)}(t), \\ &\textit{if } k = J(t), \textit{ then } U_{j(t)}(t) \leq U_{J(t)}(t). \end{aligned}$$

Proof. We can divide this proof into two cases based on which of the two arms is selected.

Case 1: let $k = j(t)$ be the arm selected. We will then assume that $L_{j(t)}(t) > L_{J(t)}(t)$ and show that this is a contradiction. By definition of the arm selection rule we know that $s_{j(t)}(t) \geq s_{J(t)}(t)$, from which we can easily deduce that $U_{j(t)}(t) > U_{J(t)}(t)$ by way of our first assumption. As a result we can see that

$$\begin{aligned}B_{j(t)}(t) &= \max_{j \neq j(t)} U_j(t) - L_{j(t)}(t) \\ &< \max_{j \neq J(t)} U_j(t) - L_{J(t)}(t) = B_{J(t)}(t).\end{aligned}$$

This inequality holds due to the fact that arm $j(t)$ must necessarily have the highest upper bound over all arms. However, this contradicts the definition of $J(t)$ and as a result it must hold that $L_{j(t)}(t) \leq L_{J(t)}(t)$.

Case 2: let $k = J(t)$ be the arm selected. The proof follows the same format as that used for $k = j(t)$. □

Corollary B2. *If arm $k = a_t$ is pulled at time t, then the minimum index is bounded above by the uncertainty of arm k, or more precisely*

$$B(t) \leq s_k(t).$$

Proof. We know that k must be restricted to the set $\{j(t), J(t)\}$ by definition. We can then consider the case that $k = j(t)$, and by Lemma B2 we know that this imposes an order on the lower bounds of each possible arm, allowing us to write

$$B(t) \leq U_{j(t)}(t) - L_{j(t)}(t) = s_{j(t)}(t)$$

from which our corollary holds. We can then easily see that a similar argument holds for $k = J(t)$ by ordering the upper bounds, again via Lemma B2. □

Lemma B3. *On event $\mathcal{E}$, for any time $t \in \{1, \ldots, T\}$, and for arm $k = a_t$ the following bound holds on the minimal gap,*

$$B(t) \leq \min(0, s_k(t) - \Delta_k) + s_k(t).$$

Proof. In order to prove this lemma we will consider a number of cases based on which of $k \in \{j(t), J(t)\}$ is selected and whether or not one or neither of these arms corresponds to the optimal arm k^*. Ultimately, this results in six cases, the first three of which we will present are based on selecting arm $k = j(t)$.

Case 1: consider $k^* = k = j(t)$. We can then see that the following sequence of inequalities holds,

$$\mu_{(2)} \overset{(a)}{\geq} \mu_{J(t)}(t) \overset{(b)}{\geq} L_{J(t)}(t) \overset{(c)}{\geq} L_{j(t)}(t) \overset{(d)}{\geq} \mu_k - s_k(t).$$

Here (b) and (d) follow directly from event $\mathcal{E}$ and (c) follows from Lemma B2. Inequality (a) follows trivially from our assumption that $k = k^*$, as a result $J(t)$ can only be as good as the 2nd-best arm. Using the definition of Δ_k and the fact that $k = k^*$, the above inequality yields

$$s_k(t) - (\mu_k - \mu_{(2)}) = s_k(t) - \Delta_k \geq 0$$

Therefore the min in the result of Lemma B3 vanishes and the result follows from Corollary B2.

Case 2: consider $k = j(t)$ and $k^* = J(t)$. We can then write

$$\begin{aligned}B(t) &= U_{j(t)}(t) - L_{J(t)}(t) \\ &\leq \mu_{j(t)}(t) + s_{j(t)}(t) - \mu_{J(t)}(t) + s_{J(t)}(t) \\ &\leq \mu_k - \mu^* + 2s_k(t)\end{aligned}$$

where the first inequality holds from event $\mathcal{E}$, and the second holds because by definition the selected arm must have higher uncertainty. We can then simplify this as

$$\begin{aligned}&= 2s_k(t) - \Delta_k \\ &\leq \min(0, s_k(t) - \Delta_k) + s_k(t),\end{aligned}$$

where the last step evokes Corollary B2.

Case 3: consider $k = j(t) \neq k^*$ and $J(t) \neq k^*$. We can then write the following sequence of inequalities,

$$\mu_{j(t)}(t) + s_{j(t)}(t) \overset{(a)}{\geq} U_{j(t)}(t) \overset{(b)}{\geq} U_{k^*}(t) \overset{(c)}{\geq} \mu^*.$$

Here (a) and (c) hold due to event $\mathcal{E}$ and (b) holds since by definition $j(t)$ has the highest upper bound other than $J(t)$, which in turn is not the optimal arm by assumption in this case. By simplifying this expression we obtain $s_k(t) - \Delta_k \geq 0$, and hence the result follows from Corollary B2 as in Case 1.

Cases 4–6: consider $k = J(t)$. The proofs for these three cases follow the same general form as the above cases and is omitted. Cases 1 through 6 cover all possible scenarios and prove Lemma B3. □

C. Modelling Bernoulli arms

Consider K Bernoulli arms, each associated with an unknown parameter $\theta_k \in \mathbb{R}$, i.e. on pulling arm k we receive reward $y = 1$ with probability θ_k. The standard, conjugate prior for such models is the Beta distribution, and we will associate each arm with a Beta(α_0, β_0) prior. The posterior for this model is then also Beta distributed such that

$$\pi_k^t(\theta_k) = \text{Beta}(\theta_k; \alpha_k(t), \beta_k(t))$$

with initial parameters corresponding to those of the prior. If arm $a_{t-1} = k$ is pulled at time $t - 1$, the posterior can be updated with

$$\alpha_k(t) = \alpha_k(t-1) + \mathbb{I}_1(y_{t-1}),$$
$$\beta_k(t) = \beta_k(t-1) + \mathbb{I}_0(y_{t-1}),$$

i.e. the parameters represent counts of successes and failures respectively. Note also that the distribution over the mean rewards is trivial in this situation as $\mu_k = \theta_k$ and as a result $\rho_k^t = \pi_k^t$.

In this model, since the posteriors are not symmetric about their mean, bounds based on the standard deviations do not necessarily represent fixed-confidence upper and lower bounds. Instead, we will use the $(1-\gamma)$th quantile and the γth quantile, respectively, where $0 < \gamma < \frac{1}{2}$ will be determined later. Let Q be the quantile function defined such that for $X \sim \rho$, $\gamma = \Pr(X \le Q(\gamma;\rho))$. We can then write upper and lower bounds

$$U_k(t) = Q(1-\gamma; \pi_k^t),$$
$$L_k(t) = Q(\gamma; \pi_k^t).$$

Corollary C3. *Consider a K-armed Bernoulli bandit problem with horizon T and let $U_k(t)$ and $L_k(t)$ be defined as above. For $\epsilon > 0$ and a quantile parameter*

$$\gamma = \exp\big\{-\tfrac{1}{2}(T + K(N_0 - 2))/H_\epsilon\big\},$$

for $N_0 = \alpha_0 + \beta_0$, then the algorithm attains simple regret satisfying

$$\Pr(R_{\Omega(T)} \le \epsilon) \ge 1 - 2KT\gamma.$$

Proof. We will first consider the lower quantile $L_k(t)$ associated with arm k at time t. Let $d(x,y)$ denote the KL-divergence between two Bernoulli random variables with parameters x and y respectively. We will also define $a = \alpha_k(t) - 1$ and $n = \alpha_k(t) + \beta_k(t) - 1$. By directly applying Lemma D2 we can bound the lower quantile with

$$L_k(t) \ge \operatorname*{arg\,min}_{x \le \frac{a+1}{n}} \Big\{x : \frac{\log(1/\gamma)}{n} \ge d\Big(\frac{a+1}{n}, x\Big)\Big\},$$

we can then use Pinsker's inequality to lower-bound the KL-divergence with a quadratic term

$$\ge \operatorname*{arg\,min}_{x \le \frac{a+1}{n}} \Big\{x : \frac{\log(1/\gamma)}{n} \ge 2\Big(\frac{a+1}{n} - x\Big)^2\Big\},$$

and finally we can lower-bound the quadratic term with the same quadratic shifted towards zero

$$\ge \operatorname*{arg\,min}_{x \le a/n} \Big\{x : \frac{\log(1/\gamma)}{n} \ge 2\Big(\frac{a}{n} - x\Big)^2\Big\} = x^-.$$

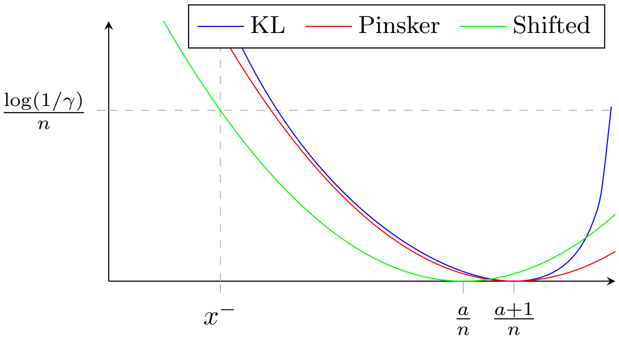

Figure 5. Illustration of the techniques used for bounding the lower quantile $L_k(t)$ from below.

Note also that we have restricted the set of possible bounds to be $x \le a/n < (a+1)/n$. Finally, using a similar application of Lemma D2 and Pinkser's inequality we can bound the upper quantile as

$$U_k(t) \le \operatorname*{arg\,max}_{x > a/n} \Big\{x : \frac{\log(1/\gamma)}{n} \ge 2\Big(\frac{a}{n} - x\Big)^2\Big\} = x^+.$$

A graphical illustration of the techniques used to obtain these bounds for the lower quantile is shown in Figure 5.

We can easily see that the possible values for both x^- and x^+ are the values of x on either side of a/n for which the quadratic term $2(a/n - x)^2$ is below $\log(1/\gamma)/n$. The bounds must then be given by the two values of x for which this quadratic term are greatest, i.e. $a/n \pm \sqrt{\log(1/\gamma)/(2n)}$. We can then define the confidence diameter as

$$s_k(t) \le x^+ - x^- = \sqrt{\frac{2\log(1/\gamma)}{\alpha_k(t) + \beta_k(t) - 1}},$$

Given the form of the Bayesian updates for this model, we know that that the parameters for each model are the success and failure counts *plus* the pseudo-counts. As a result we can obtain the arm pull counts by subtracting the pseudo-counts, i.e.

$$N_k(t) = \alpha_k(t) + \beta_k(t) - N_0$$

We can then rewrite the confidence diameter in order to define $g(N)$ as

$$s_k(t) \le \sqrt{\frac{2\log(1/\gamma)}{N_k(t) + N_0 - 1}} = g(N_k(t)).$$

The bounding function g is strictly monotonically decreasing and is invertible with

$$g^{-1}(s) = 1 - N_0 + \frac{2\log(1/\gamma)}{s^2}.$$

By setting $\sum_{k=1}^{K} g^{-1}(H_{k\epsilon}) = T - K$ and solving for γ, we arrive at the definition of γ given in the statement of this corollary. Finally, we can easily see by the definition of the quantile function that $U_k(t)$ and $L_k(t)$ bound the expected reward μ_k with probability $1 - 2\gamma$ for all k and t. These last two remarks satisfy the conditions of Theorem 1, thus completing our proof. □

D. Model-specific lemmas

Lemma D1. *Consider a normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ and $\beta \geq 0$. The probability that X is within a radius of $\beta\sigma$ from its mean can then be written as*

$$\Pr\big(|X - \mu| \leq \beta\sigma\big) \geq 1 - e^{-\beta^2/2}.$$

Proof. This proof follows from that of (Srinivas et al., 2010). Consider $Z \sim \mathcal{N}(0, 1)$. The probability that Z exceeds some positive bound $c > 0$ can be written

$$\begin{aligned}\Pr(Z > c) &= \frac{e^{-c^2/2}}{\sqrt{2\pi}} \int_c^\infty e^{(c^2 - z^2)/2}\, dz \\ &= \frac{e^{-c^2/2}}{\sqrt{2\pi}} \int_c^\infty e^{-(z-c)^2/2 - c(z-c)}\, dz \\ &\leq \frac{e^{-c^2/2}}{\sqrt{2\pi}} \int_c^\infty e^{-(z-c)^2/2}\, dz = \tfrac{1}{2} e^{-c^2/2}.\end{aligned}$$

The inequality holds due to the fact that $e^{-c(z-c)} \leq 1$ for $z \geq c$. Using a union bound we can then bound both sides as $\Pr(|Z| > c) \leq e^{-c^2/2}$. Finally, by setting $Z = (X - \mu)/\sigma$ and $c = \beta$ we obtain the bound stated above. □

Lemma D2. *This proof follows a similar structure to that of (Kaufmann et al., 2012a). Consider $X \sim \mathrm{Beta}(a, b)$ for integers a and b. Let $d(x, y)$ denote the KL-divergence between two Bernoulli random variables with parameters x and y respectively. For $\gamma < 0.5$, let $q_\gamma = Q(\gamma, \mathrm{Beta}(a, b))$ denote the lower γth quantile for X and similarly define the upper quantile as $q_{1-\gamma}$. These quantiles can then be bounded as follows:*

$$\begin{aligned} q_\gamma &\geq \operatorname*{arg\,min}_{x < \frac{a}{a+b-1}} \Big\{ x : \tfrac{\log(1/\gamma)}{a+b-1} \geq d\big(\tfrac{a}{a+b-1}, x\big) \Big\}, \\ q_{1-\gamma} &\leq \operatorname*{arg\,max}_{x > \frac{a-1}{a+b-1}} \Big\{ x : \tfrac{\log(1/\gamma)}{a+b-1} \geq d\big(\tfrac{a-1}{a+b-1}, x\big) \Big\}. \end{aligned}$$

Proof. We will first note that for a collection of $a+b-1$ standard, uniform random variables the ath order statistic has distribution $\mathrm{Beta}(a, b)$. Letting S_{nx} denote a binomial random variable with n trials and success probability x, we can think of $S_{a+b-1,x}$ as the number of uniform variates that are smaller than x.

Based on this observation we can relate the cumulative distribution of X to the number of uniform random variables that lie above X, i.e.

$$\begin{aligned}\Pr(X \leq x) &= \Pr(S_{a+b-1,x} \geq a) \\ &\leq \exp\big\{ -(a+b-1)\, d\big(\tfrac{a}{a+b-1}, x\big)\big\}.\end{aligned}$$

This last inequality holds by Sanov's theorem for $x < a/(a+b-1)$. We can then note the following sequence of implications:

$$\begin{aligned}&\exp\big\{ -(a+b-1)\, d\big(\tfrac{a}{a+b-1}, x\big)\big\} \leq \gamma \\ &\qquad \Rightarrow \Pr(X \leq x) \leq \Pr(X \leq q_\gamma) \Rightarrow x \leq q_\gamma.\end{aligned}$$

The last inequality follows from the fact that $\Pr(X \leq x)$ increases monotonically in x. As a result we can find the minimum value of x for which the first inequality holds, which leads to the lower bound given in the statement of this lemma.

Bounding the upper quantile relies on a very similar sequence of steps. We can first relate the probability that X exceeds some x to a binomial random variable

$$\begin{aligned}\Pr(X \geq x) &= \Pr(S_{a+b-1,x} \leq a-1) \\ &= \Pr(S_{a+b-1,1-x} \geq b) \\ &\leq \exp\big\{ -(a+b-1)\, d\big(\tfrac{b}{a+b-1}, 1-x\big)\big\} \\ &= \exp\big\{ -(a+b-1)\, d\big(\tfrac{a-1}{a+b-1}, x\big)\big\}.\end{aligned}$$

The final inequality follows from the fact that the KL-divergence satisfies $d(y, 1-x) = d(1-y, x)$. Note also that, as before, the inequality follows from Sanov's theorem, this time for $x > (a-1)/(a+b-1)$. Finally we can then note the following sequence of implications

$$\begin{aligned}&\exp\big\{ -(a+b-1)\, d\big(\tfrac{a-1}{a+b-1}, x\big)\big\} \leq \gamma \\ &\qquad \Rightarrow \Pr(X \geq x) \leq \Pr(X \geq q_{1-\gamma}) \Rightarrow x \geq q_{1-\gamma}.\end{aligned}$$

Here the final implication holds due to the fact that $\Pr(X \geq x)$ decreases monotonically in x. From here we can take the maximum x for which this bound holds in order to obtain the desired bound on the upper quantile. □